doi:10.2498/iti.2012.0486

Very Many Variables and Limited Numbers of Observations; The p>>n Problem in Current Statistical Applications

Johann Sölkner
Division of Livestock Sciences
University of Natural Resources and Life Sciences
Vienna, Austria
E-mail: johann.soelkner@boku.ac.at

Abstract

New technologies have led to an "explosion" of data available to document states and processes in very many fields. Tools of data mining are being used to extract relevant information. If this information is used in decision making, analytical statistics can provide formal tests comparing the outcomes of different scenarios. Statistics has traditionally dealt with limited information, both in terms of observations and numbers of variables explaining the states of these observations. Virtually all statistical hypothesis testing was developed for such scenarios, trying to make sense from limited data, often expensive to produce. Clinical trials and the steps in development of drugs before those clinical trials are a typical examples from human medicine.

Information about the genomes of individuals from DNA is becoming cheaper at an extremely fast rate. The DNA of humans and many animal species is composed of ~3.000.000.000 base pairs (nucleotides), arranged in 20-40 pairs of chromosomes. Chips extracting genetic variation from 500.000 to 2.500.000 genomic markers cost $100-300. Pharmacogenetics and –genomics refer to genetic differences in metabolic pathways which can affect individual responses to drugs, both in terms of therapeutic effect as well as adverse effects. Clearly, linking genomic information with the outcomes of a clinical trial yields a p>>n problem.

We will link genomic information from human and livestock species to phenotypes in order to elucidate developments in statistics dealing with the p>>n problem. In this context, the ability to predict outcomes is often as important as the ability to understand the cause–effect relationship. Classical statistics is augmented with data mining tools. We will learn about different types of variable selection procedures trying to extract the most important explanatory variables and we will also deal with multivariate black-box approaches. From this perspective, we will look at similar scenarios in other fields of biology, ecology and economics.